A GIS-Based, Case-Control Analysis of Cancer Incidence and Land Use Patterns

Steve Dearwent*
Department of Environmental Health Sciences, School of Public Health, University of Alabama at Birmingham, Birmingham, AL

Abstract

Geographic information systems (GIS) have been used in environmental epidemiologic studies primarily for ecologic analyses. However, many public health researchers are aware of the limitations of the ecologic study when compared with cohort and case-control designs. This paper outlines an approach to be used in a GIS-dependent, case-control analysis of cancer incidence and land use patterns. The study base consists of residents in Jefferson County, Alabama, a large metropolitan area with a population of approximately 650,000. Incident cases of three primary cancers (brain/central nervous system, non-Hodgkin's lymphoma, and pancreas) are identified through the Alabama Statewide Cancer Registry. A static residential requirement of five years is imposed on study subjects to estimate a minimal latency period for neoplastic development and control for population mobility. Georeferencing of cases and controls is anticipated to be highly accurate due to linkage with tabular data and related digitized parcel coverages maintained by the county. As with many GIS-based health studies, distance is a surrogate for exposure and is assessed using buffers generated around residential parcels. Land use characteristics are defined for every parcel in the county (approximately 290,000) and are divided into 16 classes ranging from agriculture and low-density residential to heavy industrial and resource extraction (mining). This study should describe the spatial distribution of these particular cancers in a major metropolitan area as well as address the potential relationships between environmental determinants and disease incidence.

Keywords: cancer, incidence, land use

Introduction

Cancer is a multifactorial disease frequently having etiologies of both environmental and genetic influence. Because of the diverse nature of cancer and the variability of anatomical characteristics exhibited by this disease, health researchers use many scales of analysis. These differing scales range from the study of the disease at molecular and cellular levels, to individual cases and large, population-based analyses. This variety of approaches has been beneficial in understanding biological processes, risk factors, and prevention efficacy as it relates to cancer morbidity and mortality.

Cancer has been recognized as a valuable indicator for environmentally related health effects because there is a definable endpoint (1). Cancers affecting many anatomical sites including bladder, blood, brain, kidney, liver, lung, prostate, and skin have been associated with exposure to synthetic chemicals in the occupational setting (2). It

* Steve M Dearwent, UAB School of Public Health, Dept. of Environmental Health Sciences, Rm. 317, Birmingham, AL 35294-0022 USA; (p) 205-934-6080; (f) 205-975-6341; E-mail: cryptcl@wwisp.com

is probable that a portion of the cancer burden also results from environmental (nonoccupational) exposures. These may include effects from natural substances (sunlight, radon), man-made influences (organic products of incomplete combustion), or a combination of both (asbestos, metals, nitrates, fluorides, exogenous hormones).

Environmental pollutants and resulting adverse health effects have an inherent spatial relationship. The distance from a contaminant source to a given population can influence the magnitude of exposure. Therefore, one may infer that proximity to a source may be a good predictor of the extent of adverse health effects attributable to that source. A geographic information system (GIS) can be used to organize and analyze data in studies designed to consider distance and locational attributes.

Historically, studies using GIS have been descriptive in nature. They have also tended to aggregate exposure/outcome into areas or groups (the ecologic analysis). From an epidemiological perspective, case-control and cohort designs hold more promise in quantifying associations between exposure and disease. Therefore, researchers using GIS should strive to incorporate location-specific measures for both exposure and disease, avoiding data aggregation techniques. Using location-specific measures increases study precision and validity. Accurate georeferencing of study subjects increases precision by reducing random error. Validity is improved with accurate exposure estimation because it increases the chance for correctly assessing cases or controls (minimizing nondifferential misclassification).

The purpose of this study is to describe the spatial variation of cancer incidence in Jefferson County, Alabama, particularly as it relates to land use. It should be emphasized that the methods for defining disease incidence and environmental determinants in this study are not based on an aggregate model. This manuscript describes the data sets, study design, and rationale for research. Analysis is not complete so results are not presented.

Data Sources/Descriptions

The three primary data sets being used in this study detail cancer incidence, residential parcel history, and land use in Jefferson County. The sources for this information are described below. The databases for parcel history and land use are already spatially referenced and accessible via the Jefferson County Information Services network. The data set for cancer incidence is spatially referenced through matching to the county's master address database and subsequent linkage with digitized parcel maps.

Cancer incidence data for Jefferson County are available from the Alabama Statewide Cancer Registry (ASCR). The ASCR began data collection on January 1, 1996. This data set is anticipated to be particularly complete for the study area because all of the hospital-based registries in Jefferson County providing data to the ASCR existed prior to the beginning of statewide data collection. There are approximately 500 combined cases for the cancers (brain/central nervous system, non-Hodgkin's lymphoma, pancreas) and time period (1996–1997) under analysis. Case totals for each category of cancer examined in this study are documented in Table 1.

The Office of Stormwater Management (OSWM) maintains the land use database. The OSWM is a nonprofit public entity that deals with environmental compliance issues pertaining to stormwater discharge in Jefferson County. It was created in response to the National Pollution Discharge Elimination System (NPDES). The stormwater

Table 1 Cancer Cases by Anatomical Site, Jefferson County, AL, 1996–1997

	Number of Cancer Cases	
Anatomical Site	**1996**	**1997**
Brain/central nervous system	46	34
Non-Hodgkin's lymphoma	129	118
Pancreas	80	84

coverage used in this study was developed over a span of four years (1991–1994) by Walter Schoel Engineering. Aerial photography and field-verified land use maps were the primary sources for creating the coverage (3). The series of aerial photographs used in this project were taken in 1990 for over 1,300 1-mile sections in the county. For regions of the county experiencing heavy growth rates, personnel were sent into the field to visually verify documented patterns. The coverage classifies every parcel in the county into one of 16 categories of land use. The magnitude of this database is immense, considering that Jefferson County is an area of over 1,120 square miles and there are approximately 290,000 individual parcels of land in the county ranging in size from small residential plots (fractions of an acre) to large commercial and government-owned properties (multiple acres/square miles). The NPDES data classify every parcel according to the scheme outlined in Table 2.

The Jefferson County Tax Assessor database is used in the analysis for many functions. This source details parcel information for the entire county and assists in enumerating the study base, georeferencing all study participants, and obtaining residential parcel characteristics (zoning, length of ownership, property value). Specific study restrictions have been applied during the query of the tax assessor database to identify all "eligible" parcels within the county. These parameters and the rationale for imposing them are discussed subsequently.

Methods

The data described above are being used in a cumulative incidence, case-control analysis of cancer and land use patterns in Jefferson County. Cases are defined as primary cancers occurring in Jefferson County that are identified through the ASCR for the period of 1996 through 1997.

Study restrictions insure that cases are derived from the same cohort (the study base) out of which controls are selected (4). These parameters are applied to the parcel of land where subjects reside. Eligible parcels must meet the following restrictions:

- Should lie completely within the boundaries of Jefferson County
- Should be zoned for residential use
- Cannot have a deed date (transaction) during the period 1992–1997
- Must have homestead status

For obvious reasons, a study subject's residential parcel must fall within the county boundaries. If this condition is not met, then that individual is not considered to be a

Table 2 Land Use Categories, Jefferson County, AL, 1991

Database Coding	Field Description	Number of Polygons[a]	Percentage of Area within Jefferson County
AG	Agriculture	960	6.3%
CH	Heavy commercial	117	0.6%
CL	Light commercial	661	0.3%
CU	Urbanized commercial	252	0.7%
HW	Highway	9	1.3%
IH	Heavy industrial	342	1.7%
IL	Light industrial	617	1.2%
INST	Institutional (schools, churches)	1,247	1.3%
OS	Open space (parks, recreational areas, greenways)	142	1.0%
RE	Resource extraction (mining)	457	1.9%
RH	High-density residential	864	3.4%
RL	Low-density residential	1,137	4.9%
RM	Moderate-density residential	1,091	9.2%
RT	Mobile home parks	128	0.1%
U	Undeveloped	1,698	64.6%
W	Water	398	1.4%

[a] The number of polygons per land use category does not correspond with the number of parcels for each respective group. Adjoining parcels with the same land use coding have been concatenated into one polygon using the dissolve function in ARC/INFO.

resident of Jefferson County. Study subjects include only Jefferson County residents because the tax assessor database, digitized parcel maps, and land use coverage are all limited to this region.

All participants must live on a parcel zoned for residential use. This restriction is imposed to enumerate a control population more precisely. By eliminating all parcels of land used for commercial, industrial, government, and other nonresidential purposes, the remaining set should constitute a viable group of parcels where people actually live.

Eligible parcels cannot have a deed date between 1992 and 1997. Deed dates within the tax assessor database indicate a parcel transaction. By imposing this restriction, all members of the study base should have lived at their current residence for a minimum of five years. The five-year period is chosen arbitrarily but provides an estimate for static residential populations. This residential exclusion period serves many purposes. It provides more plausibility to the study design by establishing a minimal latency period for initiation and progression of neoplastic growth to diagnostic levels. Five years is an extremely short latency period, but extending this to 10, 15, or even 20 years would severely compromise the size of the study population. The five-year exclusion also assists in controlling for population mobility, an important consideration when studying locational attributes of disease status within urban areas. This is particularly true for urban residents in the US because they exhibit some of the highest levels of mobility for

any industrialized nation. The main limitation in imposing the five-year residential requirement is that it will decrease study precision (power) by eliminating cases that do not meet this parameter.

The exclusion of parcels without homestead status will eliminate potential cohort members who rent their residence. The exclusion of renters is necessary because it is virtually impossible to follow their residential history with the county records used in this study. Renters are a more mobile population and many would probably not meet the five-year static residential requirement. Imposing this parameter increases study validity because it strengthens the definition of the study base by minimizing selection bias. However, it also decreases precision because some cases will be excluded from the analysis.

The time frame under analysis in this study is used for many reasons. Case information from the ASCR is available only for this span. Also, this period corresponds well with potential exposure to the 1991 land use coverage combined with the five-year static residential requirement. In other words, all study subjects identified during the 1996–1997 time frame must have lived at their current residence since 1991/1992, approximately the same period during which the land use audit (exposure assessment) was conducted.

The processes for georeferencing cases and controls differ slightly. They are both linked to their residential parcel via digitized parcel maps. However, controls are initially defined by querying the tax assessor database for "eligible" parcels, while cases are provided by ASCR. Because all eligible controls are identified and accurately georeferenced by querying the tax assessor database, there is no need for matching address fields. If a parcel meets all the study restrictions, then it is selected along with the spatial references documented in the digitized parcel coverage. Cases, however, are georeferenced by matching street addresses documented in the ASCR database to the county's master address database, with subsequent linkage to digitized parcel maps. The matching of text-based address fields between databases is more cumbersome and problematic.

Once all study subjects are georeferenced, exposure is assessed by generating concentric buffers of predefined size around each parcel polygon label point and aggregating land use characteristics found therein. Because distance is a surrogate for exposure, varying buffer sizes will assist in dose-response and trend analyses. Buffering around points instead of parcel boundaries insures that the area encompassed by buffers using the same predefined diameter will not vary. If boundaries (polygons) of residential parcels were used to determine buffering dimensions, the buffered regions would vary in size corresponding with parcel size. They would also take on heterogeneous shapes. These factors may produce "spatial" confounding.

Discussion

Health outcome data are often georeferenced to areal units such as state, county, municipality, zip code, or census tract. They are infrequently assigned a point value even though this provides a much more accurate, non-aggregate, locational description. This practice stems from the fact that most health outcome datasets include either information on an areal unit or have a data element that can be easily related to a region.

The method of georeferencing used in this analysis is anticipated to be highly

accurate compared with typical procedures involving linkage with street address ranges. The US Census Bureau's TIGER/Line files provide a common means for matching a list of addresses to street segments and their respective address ranges. However, the use of address ranges can be problematic. Most address matching programs allocate addresses at evenly spaced intervals along a street line segment, recessed off the street by a predefined value. An apartment complex located at one end of a street may account for the majority of addresses on that segment, yet addresses will be allocated at equal intervals along the entire path. With the method for georeferencing used in this analysis, study subjects will be linked directly to the parcel of land where their residence is located. This will eliminate the erroneous assumption of evenly spaced, single-dwelling residences.

Conclusion

Geographic information systems are being integrated into many of today's information management sectors. GIS is already an important component of earth sciences. This growth will eventually have a substantial impact on the collection, management, and analysis of health outcome data. GIS provides environmental health researchers with the ability to combine data from population-based cancer registries and environmental hazard assessments. For cancers in which environmental exposures are potential risk factors, it will mature as a more useful analytical tool. Public health is beginning to witness the use of GIS in many facets, although this is not always published in the standard epidemiologic and environmental health literature (5,6,7). This growth should continue, as information technologies become an increasingly important part of our society.

Acknowledgments

This research is supported in part by grant CA47888 from the National Cancer Institute. The author is also indebted to Jefferson County Information Services for providing GIS facilities and technical guidance.

References

1. Sherman JD. 1994. *Chemical exposure and disease.* New Jersey: Princeton Scientific Publishing Co.
2. Doll R, Peto R. 1981. The causes of cancer: Quantitative estimates of avoidable risks of cancer in the United States. *Journal of the National Cancer Institute* 66:1191–1308.
3. Haynes D. 1998. Personal communication. Walter Shoel Engineering.
4. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. 1992. Selection of controls in case-control studies. *American Journal of Epidemiology* 135:1019–28.
5. McGarigle B. 1998. GIS takes on TB. *Government Technology* 6:19–20.
6. Nygeres T, et al. 1997. Geographic information systems for risk evaluation: Perspectives on applications to environmental health. *Cartography and Geographic Information Systems* 3:123–44.
7. Zhou Y, et al. 1996. GIS-based network models of Schistosomiasis infection. *Geographic Information Sciences* 2:51–7.